eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers

Year: 2022 | Citations: 963 | Authors: Y. Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song

Abstract

Large-scale diffusion-based generative models have led to breakthroughs in text-conditioned high-resolution image synthesis. Starting from random noise, such text-to-image diffusion models gradually synthesize images in an iterative fashion while conditioning on text prompts. We find that their synthesis behavior qualitatively changes throughout this process: Early in sampling, generation strongly relies on the text prompt to generate text-aligned content, while later, the text conditioning is almost entirely ignored. This suggests that sharing model parameters throughout the entire generation process may not be ideal. Therefore, in contrast to existing works, we propose to train an ensemble of text-to-image diffusion models specialized for different synthesis stages. To maintain training efficiency, we initially train a single model, which is then split into specialized models that are trained for the specific stages of the iterative generation process. Our ensemble of diffusion models, called eDiff-I, results in improved text alignment while maintaining the same inference computation cost and preserving high visual quality, outperforming previous large-scale text-to-image diffusion models on the standard benchmark. In addition, we train our model to exploit a variety of embeddings for conditioning, including the T5 text, CLIP text, and CLIP image embeddings. We show that these different embeddings lead to different behaviors. Notably, the CLIP image embedding allows an intuitive way of transferring the style of a reference image to the target text-to-image output. Lastly, we show a technique that enables eDiff-I's"paint-with-words"capability. A user can select the word in the input text and paint it in a canvas to control the output, which is very handy for crafting the desired image in mind. The project page is available at https://deepimagination.cc/eDiff-I/